Claims

1. A method of cooperatively load-balancing a cluster of server computer systems for servicing client requests issued with respect to a plurality of client computer systems, said method comprising the steps of:

a) selecting, by a client computer system, a target server computer system from said cluster of server computer systems to service a particular client request using available accumulated selection basis data;

b) evaluating, by said target server computer system, said particular client request to responsively provide instance selection basis data dynamically dependent on the configuration of said target server computer and said particular client request; and

c) incorporating said instance selection basis data into said available accumulated selection basis data to affect the subsequent selection of said target computer system with respect to a subsequent instance of said particular client request.

2. The method of Claim 1 wherein said instance selection basis data includes a representation of a dynamically determined performance level of said target server computer system and wherein said available accumulated selection basis data incorporates said instance selection basis data with identifications of said target server computer and said particular client request.

3. The method of Claim 2 wherein said instance selection basis data includes

a representation of a policy evaluation of said particular client request relative to said target server computer system.

4. The method of Claim 1 wherein said instance selection basis data includes a load value and a selection weighting value, wherein said load value represents a dynamically determined performance level of said target server computer system and said selection weighting value represents a policy evaluation of said particular client request relative to said target server computer system and wherein said available accumulated selection basis data incorporates said instance selection basis data with identifications of said target server computer and said particular client request.

5. The method of Claim 4 wherein said step of selecting selects said target server computer system based on predetermined selection criteria including the relative values of said load value and said selection weighting value with respect to said particular client request as recorded in said available accumulated selection basis data.

6. The method of Claim 5 wherein said instance selection basis data provides for a rejection of said particular client request and wherein said step of selecting includes selecting an alternate server computer system from said cluster of server computer systems as said target server system to service said particular client request based on said available accumulated selection basis data.

7. A method of load-balancing a cluster of server computer systems in the cooperative providing of a network service, said method comprising the steps of:

a) selecting, by each of a plurality of host computers, server computers within a computer cluster to which to issue respective service requests;

b) responding, by a corresponding one of said plurality of host computers, to the rejection of a predetermined service request by selecting a different server computer to which to issue said predetermined service request;

c) receiving, in regard to said respective service requests by the respective ones of said plurality of host computers, load and weight information from the respective server computers; and

d) evaluating, by each of said plurality of host computers, the respective load and weight information received with respect to server computers of said computer cluster as a basis for a subsequent performance of said step of selecting.

8. The method of Claim 7 further comprising the step of determining said weight information by each of said server computers with respect to each service request received, said weight information being determined from a predefined policy association between a received service request and the identity of the one of said server computers that receives the service request.

9. The method of Claim 8 further comprising the step of distributing initial information by said cluster of server computers to said host computers, said initial information providing selection lists of said server computers to said host computers.

10. The method of Claim 9 wherein said load information is representative of a plurality of load factors including network loading and processor loading.

11. The method of Claim 10 wherein said load information is representative of the processing of a current set of service requests including a plurality of processor functions.

12. The method of Claim 11 wherein said load information includes one or more load values representing processing functions internal to a server computer.

13. A server cluster operated to provide a load-balanced network service, said server cluster comprising:

a) a plurality of server computers individually responsive to service requests to perform corresponding processing services, wherein said server computers are operative to initially respond to said service requests to provide load and weight values, wherein said load and weight values represent the current operating load a policy-based priority level of a respective server computer relative to a particular service request; and

b) a host computer system operative to autonomously issue said service requests respectively to said plurality of server computers, said host computer system further operative to select a target server computer from said plurality of server computers to receive an instance of said particular service request based on said load and weight values.

14. The server cluster of Claim 13 wherein said host computer is operative to collect said load and weight values from said plurality of server computers in connection with the issuance of respective service requests to said plurality of server computers and wherein the selection of said target server computer is based on the relative temporal age of said load and weight values.

15. The server cluster of Claim 14 wherein each of said plurality of server computers include a policy data set store that provides for the storage of a distinct server configuration and wherein said load and weight values are dynamically determined by said plurality of server computers in response to said service requests based on said distinct server configurations of said plurality of server computers.

16. The server cluster of Claim 15 wherein said distinct server configurations include the distinct identities of said plurality of server computers.

17. The server cluster of Claim 16 wherein said distinct server configurations include distinct policy data relative to said service requests, wherein said host computer system is operative to collect, relative to respective said service requests, and provide attribute data to said plurality of server computers, and wherein said server computers evaluate said attribute data in conjunction with said distinct policy data to determine said weight values.

18. The server cluster of Claim 17 wherein said plurality of server computers implement a security processing service, wherein said host computer system is

operative to selectively route network transported data through said server computers dependent on said service requests as evaluated by said plurality of server computers.

19. The server cluster of Claim 18 said host computer is operative to initiate respective data transfer transactions for each of said service requests, wherein the default routing of each said data transfer transaction initially provides for the transfer of corresponding ones of said service requests to respective ones of said plurality of server computers, and wherein said respective ones of said plurality of server computers determine whether the subsequent routing of network data within said respective data transfer transactions includes routing said network data within said respective data transfer transactions through said plurality of server computers.

20. A computer system providing, on behalf of client computer systems, a network service through a scalable cluster of server computer systems, said system comprising:

a) a plurality of server computers coupled to provide a defined service, wherein a server computer of said plurality provides a response, including load information, in acknowledgment of a predetermined service request issued to said server computer system, said response selectively indicating nonacceptance of said predetermined service request; and

b) a client computer system having an identification list of said plurality of server computer systems, said client computer system being operative to autonomously select a first server computer system from said identification list to

which to issue said predetermined service request, wherein said client computer system is reactive to said response, on indicated nonacceptance of said predetermined service request, to autonomously select a second server computer system from said identification list to which to issue said predetermined service request, and wherein said client computer system is responsive to said load information of said response in subsequently autonomously selecting said first and second server computer systems.

21. The computer system of Claim 20 wherein said response further includes weight information and wherein said client computer system evaluates the combination of said load and weight information in autonomously selecting server computer systems from said identification list.

22. The computer system of Claim 21 wherein said plurality of server computer systems include respective policy engines and wherein said weight information reflects an association between a server computer policy role and said predetermined service request.

23. The computer system of Claim 22 wherein said predetermined service request includes predetermined client process attribute information and wherein said respective policy engines are responsive to said predetermined client process attribute information in determining said server computer policy role relative to said predetermined service request.

24. The computer system of Claim 23 wherein said load information includes a value representing network and server processor performance.

25. A method of dynamically managing the distribution of client requests to a plurality of server computer systems providing a network service, each of said server computer systems being discretely configured to respond to client requests, said method comprising the steps of:

a) processing client requests to select for a particular client request a particular server computer system of said plurality of server computer systems to service said particular client request, wherein the selection of said particular server computer system is dependent on the evaluation of accumulated selection qualification information;

b) forwarding said particular client request to said particular server computer system; and

c) receiving from said particular server computer system with respect to said particular client request instance selection qualification information discretely determined by said particular server computer system with respect to said particular client request, wherein said instance selection qualification information is incorporated into said accumulated selection qualification information.

26. The method of Claim 25 wherein said processing step dynamically evaluates said particular client request with respect to said accumulated selection qualification information to identify said particular server computer system as a best choice of said plurality of server computer systems for selection.

27. The method of Claim 26 further comprising the step of evaluating by said particular server computer system, subject to the discrete configuration of said particular server computer system, said particular client request to provide said instance selection qualification information.

28. The method of Claim 27 wherein said step of evaluating provides for the dynamic generation of said instance selection qualification information including a load value reflective of the performance capability of said particular server computer system.

29. The method of Claim 28 wherein said instance selection qualification information includes a relative prioritization of said particular client request with respect to said particular server computer system.

30. The method of Claim 29 wherein said client requests are issued with respect to client computer systems, wherein said particular client request includes attributes descriptive of a particular client computer system that issued said particular client request, and wherein said relative prioritization reflects the evaluation of said attributes with respect to said particular server computer system.

31. A method of distributing computational load over a plurality of server systems provided to support execution of a data processing service on behalf of a plurality of client systems, wherein the computational load is generated in response to client requests issued through a plurality of client processes, said method comprising the steps of:

a) first processing a particular client request to associate attribute data from a respective client process of said plurality of client processes with said particular client request;

b) selecting, for said particular client request, a particular target server system from among said plurality of server systems by matching said particular client request against accumulated selection information to identify said particular target server system;

c) second processing said particular client request, including said attribute data, by said particular target server system to dynamically generate instance selection information including a load value for said particular target server system and reflective of the combination of said particular client request and said particular target server system; and

d) incorporating said instance selection information into said accumulated selection information for subsequent use in said step of selecting.

32. The method of Claim 31 wherein said instance selection information includes a relative weighting value reflective of the combination of said particular client request and said particular target server system and wherein said step of selecting matches said particular client request, including said attribute data, against corresponding data of said accumulated selection information to choose said particular target server system based on a best corresponding combination of relative weighting value and load value.

33. The method of Claim 32 wherein said step of selecting includes a step of aging said accumulated selection information.

34. The method of Claim 33 further comprising the steps of:

a) first providing, through a host process, said particular client request, including attribute data, to said particular target server system; and

b) receiving by said host process, a particular target server response including said instance selection information;

c) determining, by said host process from said particular target server response, whether to select an alternate target server system;

d) reselecting, for said particular client request, a secondary target server system from among said plurality of server systems by matching said particular client request against said accumulated selection information, including said instance selection information received from said particular target server response to identify said secondary target server system; and

e) second providing, through said host process, said particular client request, including attribute data, to said alternate target server system.

35. The method of Claim 34 wherein said host process is executed on a client computer system.

36. The method of Claim 35 wherein said host process is executed on a gateway computer system coupleable through a communications network with a plurality of client computer systems.